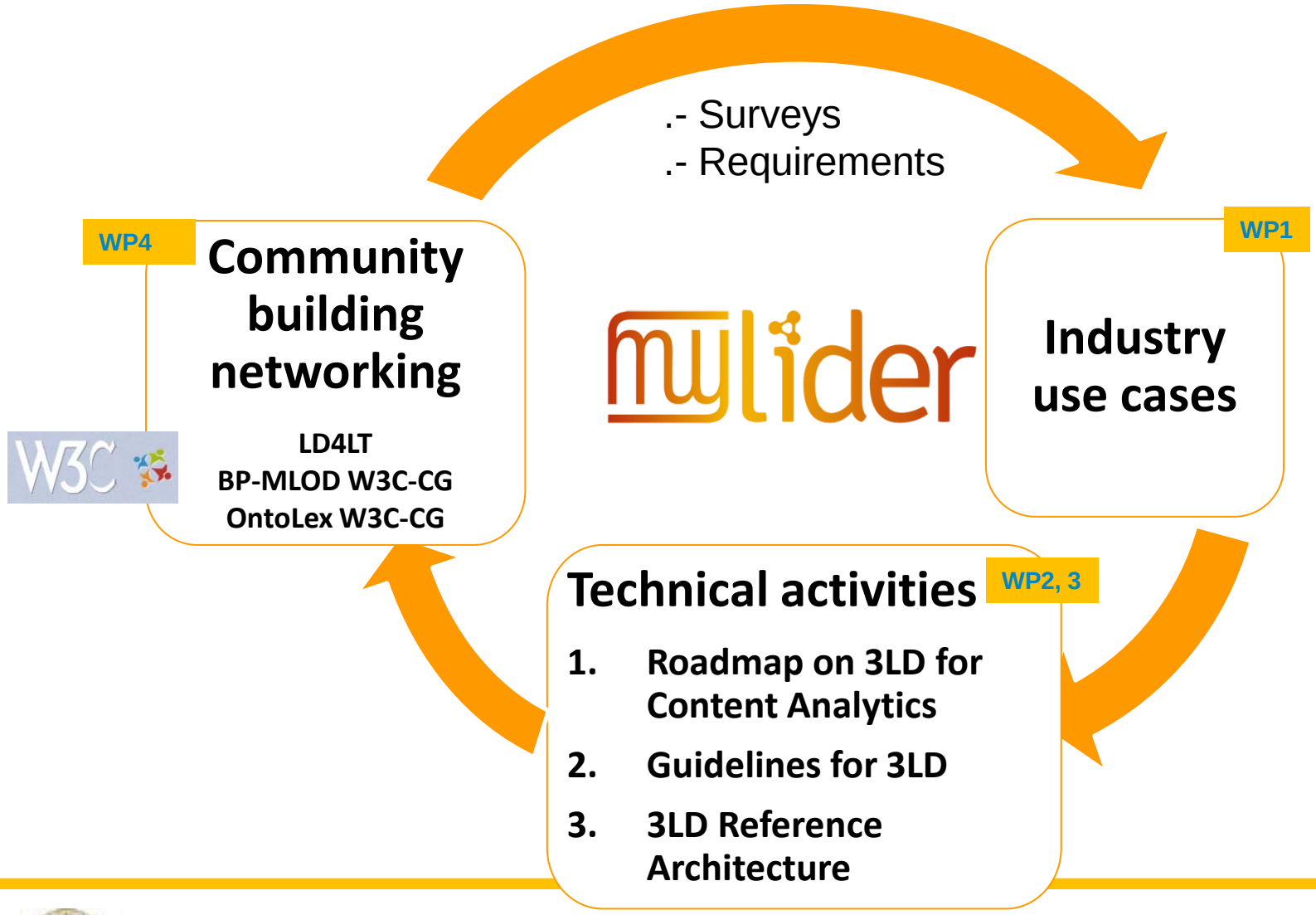
Linked Data for sharing, discovery and re-use of Language Resources at a Web scale

A. Gómez-Pérez

Universidad Politécnica de Madrid

asun@fi.upm.es

Acknowledgement: *J. Gracia , V. Rodriguez, P. Cimiano*



http://www.w3.org/community/ld4lt/
http://www.w3.org/community/bpmlod/

Lack of interoperability: semantic and technical

- Ecosystem of
  - Open and Closed resources
  - Silos of LRs
  - Complementary resources
    - Lexicon, Corpora, Dictionaries, Grammars, ….
  - Heterogeneous formats
    - E.g, for Lexicons: Lexinfo, LMF, LIR, Lemon, …
  - Several repositories with different metadata and schemas
  - Many APIs and services for querying

Discovery and reuse LR in third party applications is hard, manual and time consuming

Use cases for LR Discovery

- Language metadata content
  - Give me bilingual dictionaries in Spanish, German , that accounts for grammatical number and gender with Creative Common licenses
- Language Resources content
  - Give me all occurrences in corpora of the token “bank” disambiguated as the WorNet synset http://wordnet-rdf.princeton.edu/wn31/108437235-n
- Language Services
  - Give me all RESTfull services that can extract terms from text in Latvian.

mylider

Wikcionario
[wik.sjo'na.rjo]
El diccionario en castellano de contenido libre

- "Red"
- Pronunciation: [red]
- Grammar category: sustantivo femenino
- Singular: "red"
- Plural: "redes"

Real Academia Española

- "Red"
- Etimologiy Del latin "rete"
- Gender: "f"
- Definition.: "Conjunto de ordenadores o de equipos informáticos conectados entre sí…."

Wikilengua del español

- "Red"
- Norm: UNE 21302-131
- English: network
- German: Netzwerk

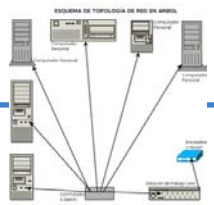
Complementary but not connected

wR

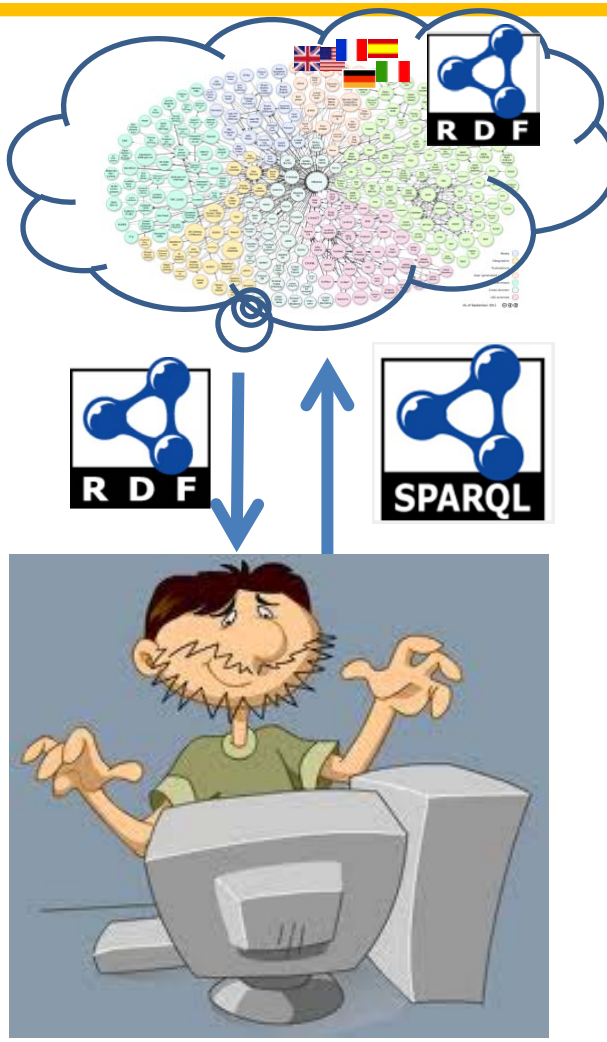
- "Red"
- Sinonyms: "sistema", "malla"," distribución"

Wikipedia

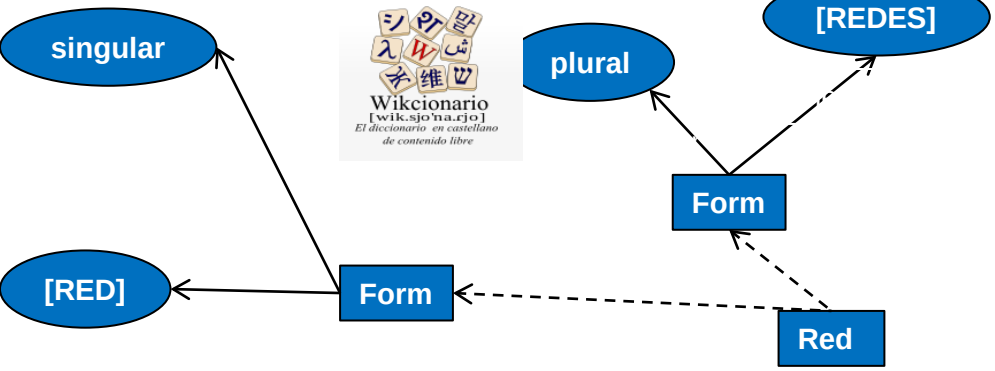
- "Red_de_computadores"
- Category: redes informáticas
- Image

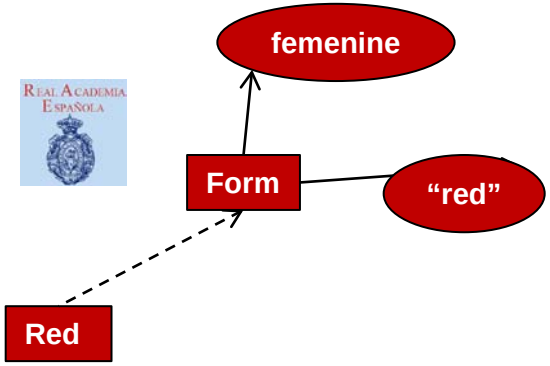


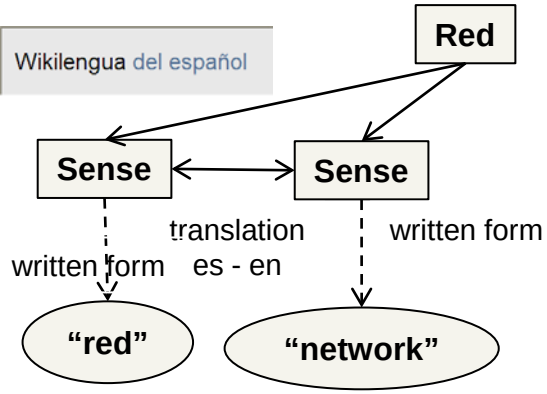
Linked Data allows uniform access to Language Resources and Services

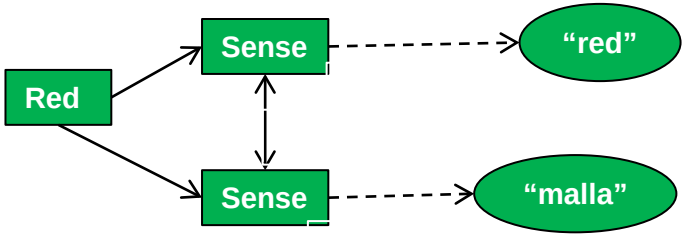


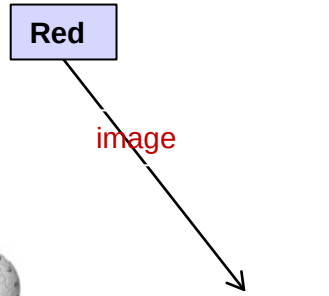
LD allows linguistic data integration



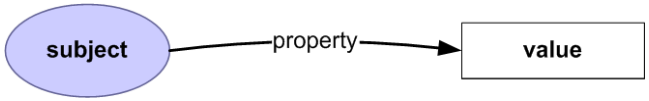








RDF(S) models



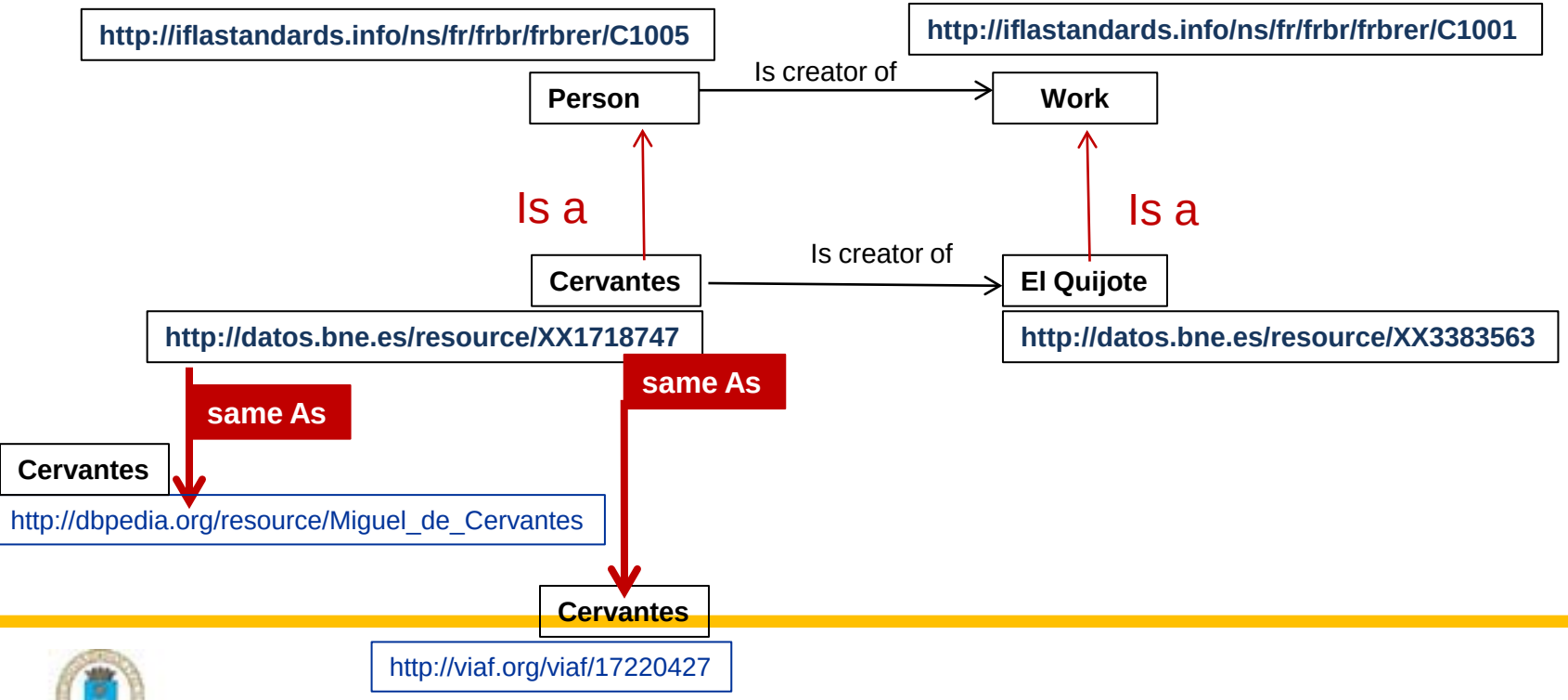
Unique identifiers: URI

identify or name a resource

Equivalence links to other datasets

Owl:same As

Data navigation



The ontology and the data

Language

http://iflastandards.info/ns/fr/frbr/frbrer/C1002

http://..../translation

Work

http://iflastandards.info/ns/fr/frbr/frbrer/C1001

http://.. /isAuthor

Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1005

http://.../date

http://.../publicaitonDate

http:///..../hasTopic

http://...../located At

Library

http://xmlns.com/foaf/0.1/Organization

Ontology

Catalán

http://datos.bne.es/resource/XX1924295

Data

http://..../translation

Don Quijote de la Mancha

http://.../is Author

Cervantes Saavedra, Miguel de

http://.../1960

http://datos.bne.es/resource/XX3383563

http://datos.bne.es/resource/XX1718747

http://.../publication Date

Tiene como materia

http://datos.bne.es/resource/bimo0002045496

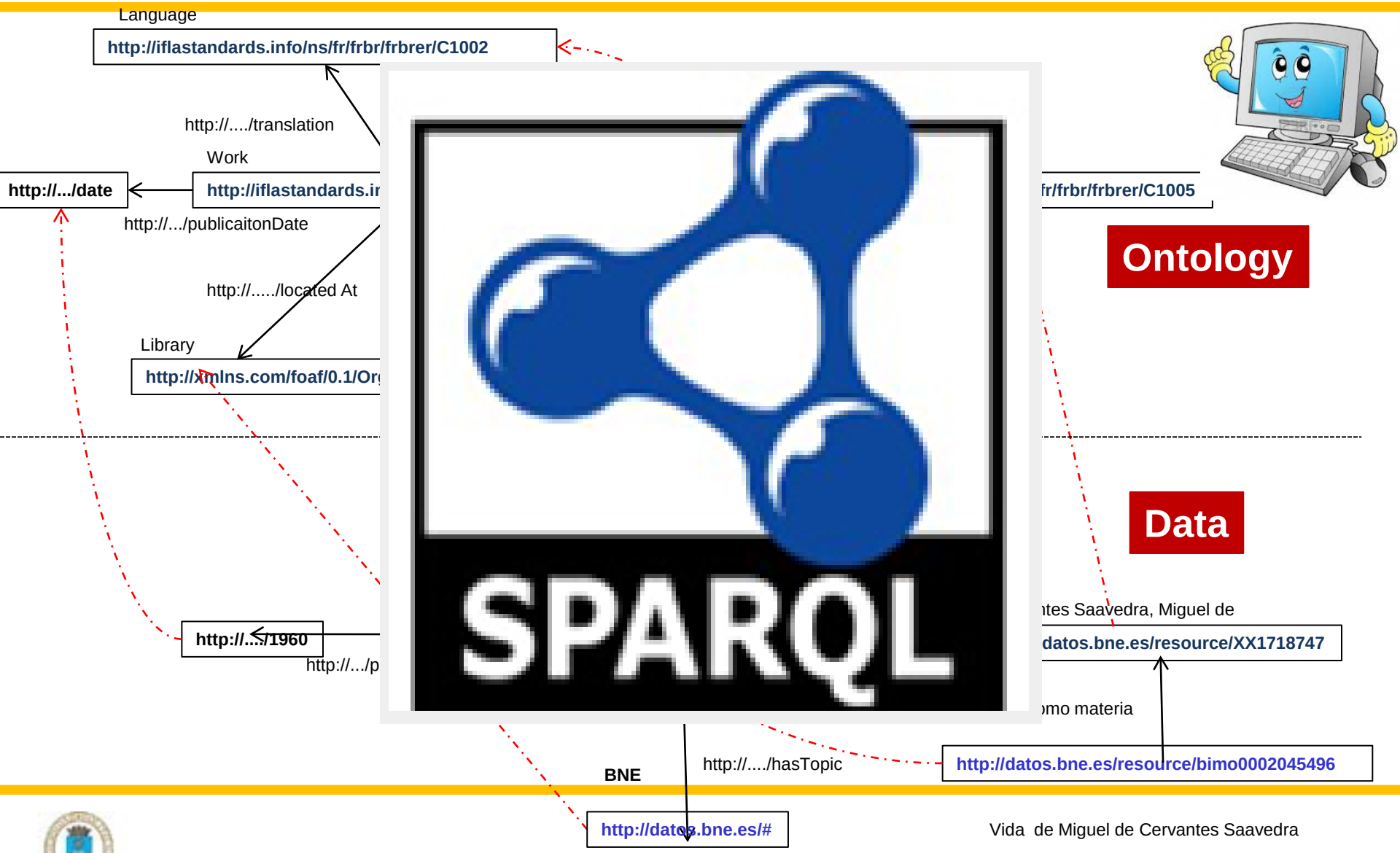
BNE

http://..../hasTopic

http://datos.bne.es/#

Vida de Miguel de Cervantes Saavedra

The query language



```
SELECT  ?l COUNT(?l) ?type
WHERE {  ?s ms:language ?l; a ?type.
FILTER((REGEX(STR(?type), "purl")))}
GROUP BY ?l ?type
```

l	[.1]	type
eng	110	ms:LexicalConceptualResource
spa	89	ms:LexicalConceptualResource
EN	49	ms:LexicalConceptualResource
fre/fra	39	ms:LexicalConceptualResource
en	36	ms:CorpusText
ger/deu	25	ms:LexicalConceptualResource
LV	24	ms:LexicalConceptualResource
DE	23	ms:LexicalConceptualResource
RU	23	ms:LexicalConceptualResource
fi	21	ms:CorpusText
FR	19	ms:LexicalConceptualResource
fi	17	ms:LexicalConceptualResource
IS	17	ms:LexicalConceptualResource
ita	16	ms:LexicalConceptualResource
et	16	ms:CorpusText
PL	15	ms:LexicalConceptualResource
fre/fra	15	ms:CorpusText

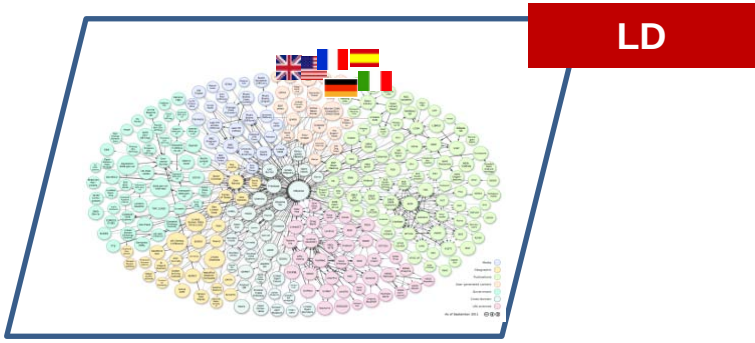
RESOURCES by LANGUAGE & TYPE?

IEA LSP

sample data (855 records)

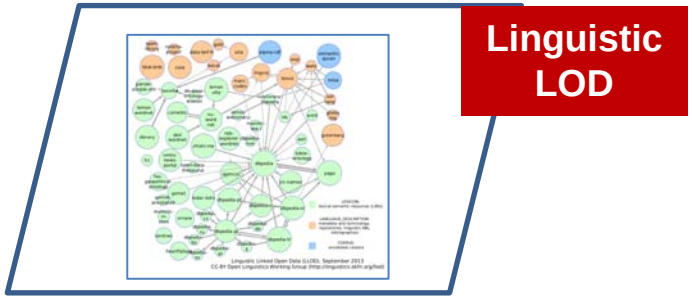
Acknowledgements: Marta Villegas and Nuria Bel

LD is increasingly multilingual



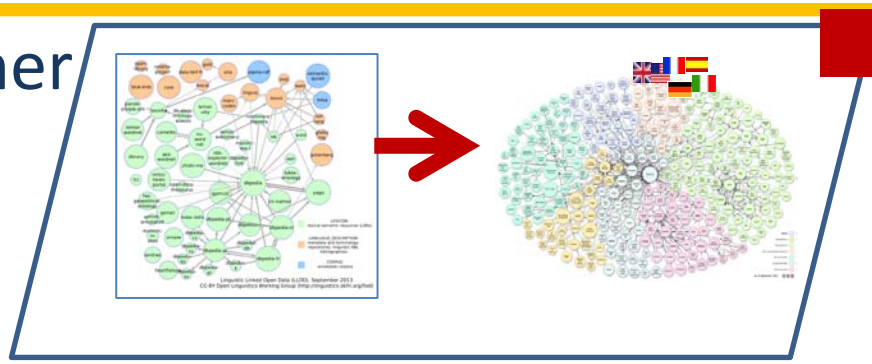
Linguistic LD

- Subset of LD focussed on LR
- Open or Close Licenses
- Resources in RDF
- Interconnected with other data
- Relevant examples of LR in RDF
  - Princeton Wordnet
  - Babelnet

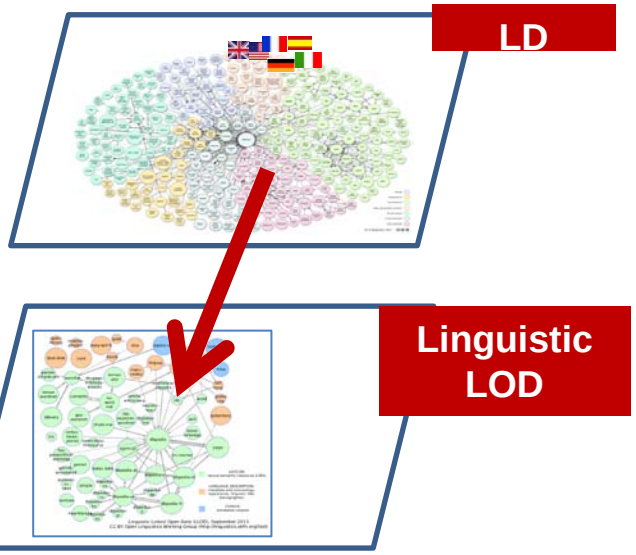


How LD and Linguistic LD is related

Is Linguistic LOD just another type of dataset to be exposed in RDF?



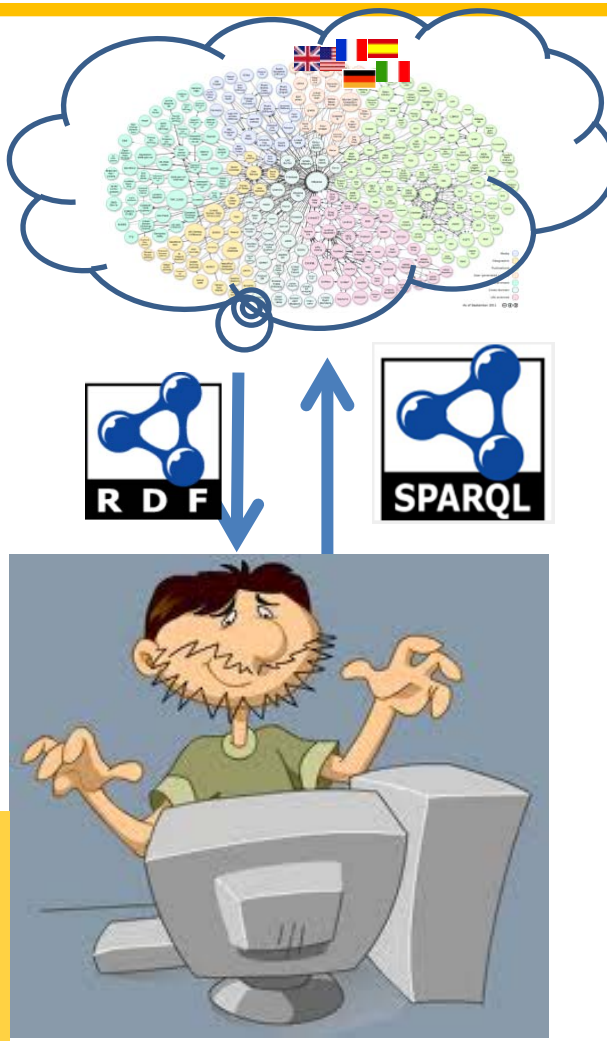
Is the role of Linguistic LOD to extend the domain LOD datasets with lexical entries?



How the Linguistic LOD generation differ from the LD generation?

Key Ingredients in Linguistic LD

1. **Agree on vocabularies** for describing
   - LR metadata
   - LR content
2. Unified and standardized language for describing resources (**RDF(S)**)
3. Unified and standardized **query language** (SPARQL)
4. Standardized non-proprietary APIs
5. Links to other resources



Additional Requirements for LRs as LD:

Keep track of the **License (open or closed)** information
Keep track of the **Provenance** of the resource
Keep track of the **use** of the resource

3LD

Linguistic Linked Licensed Data

Language resources such as:

- *Lexica*
- *Corpora*
- *Dictionaries*
- *Grammars ..*

*Using **RDF** and **standard data models** (vocabularies):*

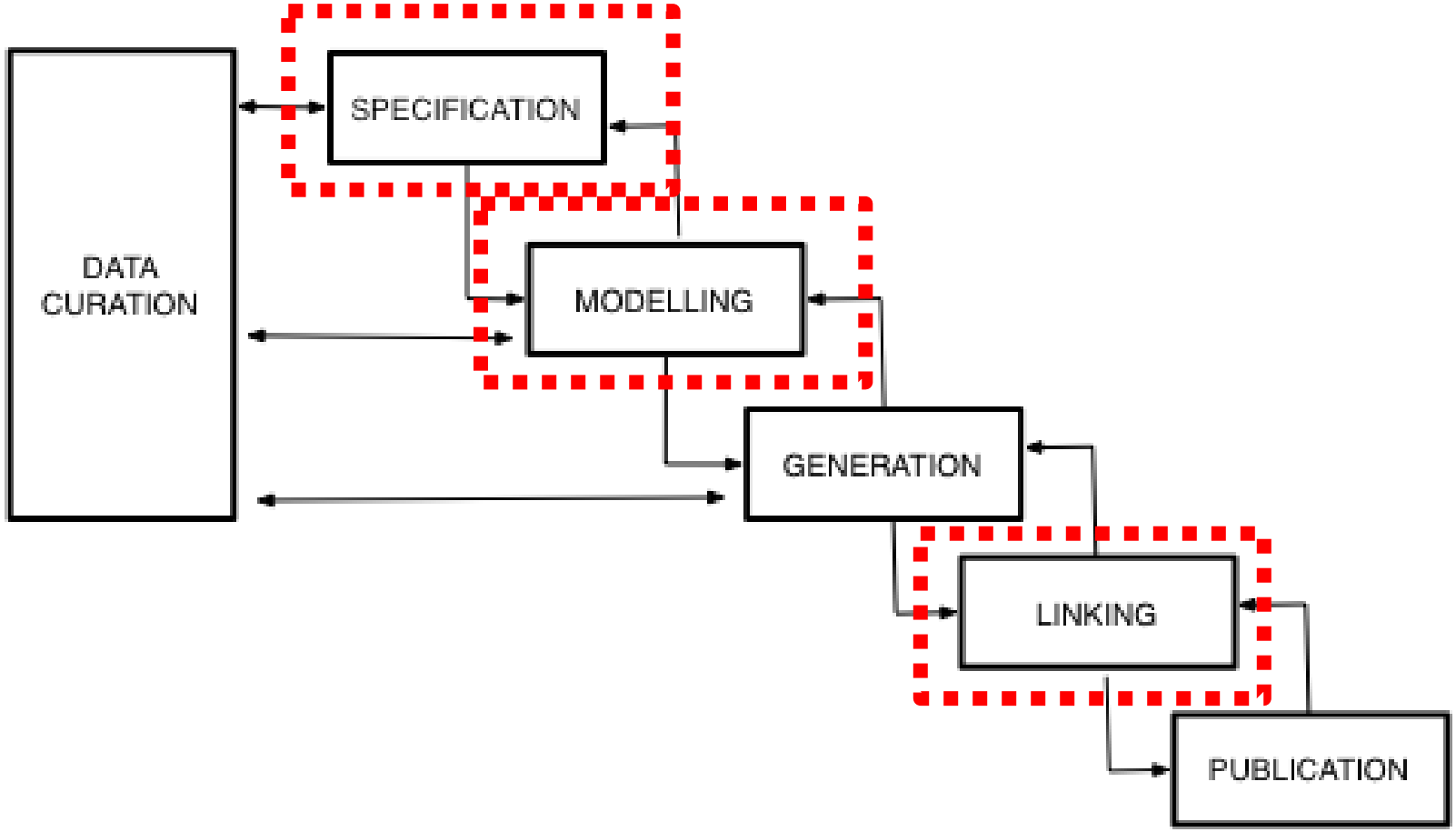
- *Lexica* lemon
- *Corpora* NIF NLP Interchange Format
- *.....*

*Published along with a **machine-readable license**.*

ODRL
Open Digital Rights Language

creative commons

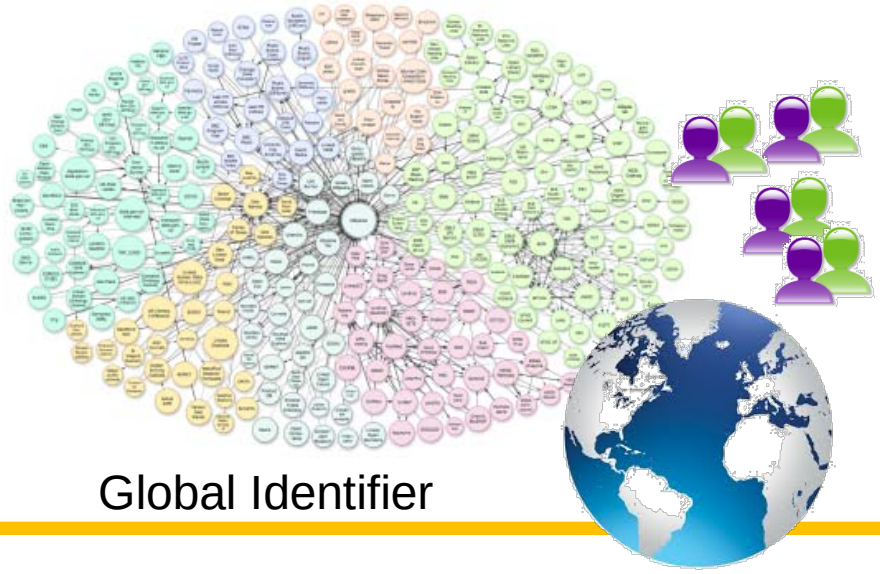
Method and lifecycle



Uniform Resource Identifier (URI)

- MANDATORY
  - To ensure **uniqueness of the resource** at the Web scale
  - To allow the Linked Data mechanisms
- Assigned by …
  - An authoritative part (E.g. ELDA)
  - Each provider can create their own URIs
  - Both can coexist

Local identifier



Global Identifier

The need of ontologies and owl: sameAs

URIs linked without ontologies

http://www.server1.org/resource/Cervantes

Same as

http://d-nb.info/gnd/11851993X

Same as

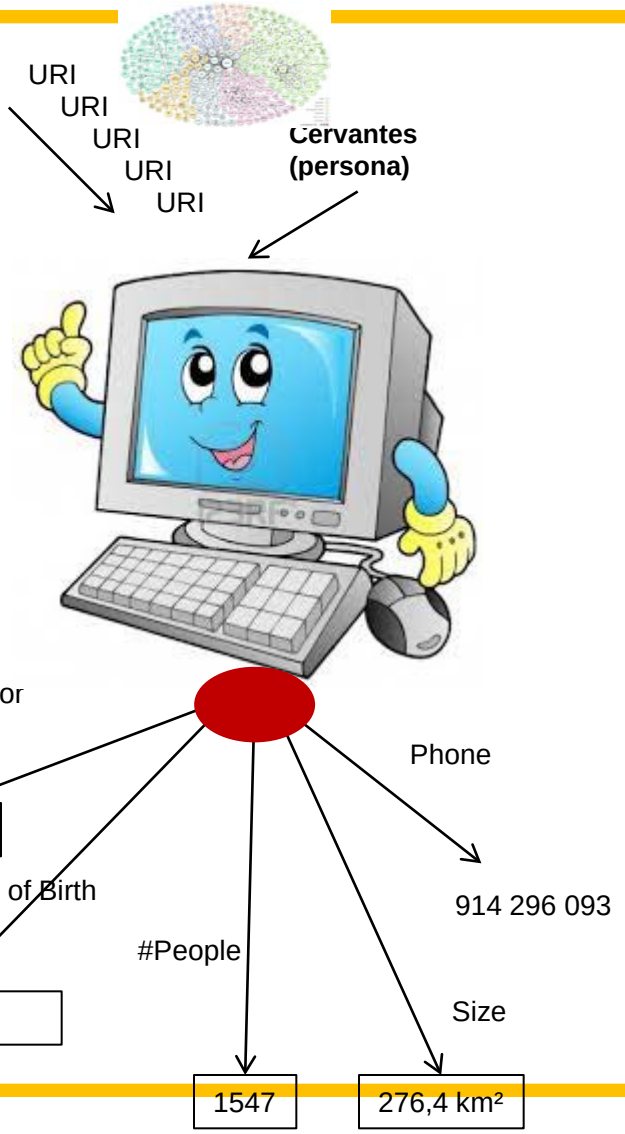
http://datos.bne.es/resource/XX1718747

Same as

http://www.server2.es/resource/Cervantes

Same as

http://geo.linkeddata.es/page/resource/Municipio/Cervantes



URIS, ontologies and owl:sameAs

rdf:type → http://..../Retaurant

http://www.server1.org/resource/Cervantes

URI
URI
URI
URI
URI

Cervantes (Person)

rdf:type → http://schema.org/Person

http://d-nb.info/gnd/11851993X

EquivalentClass

Same as

rdf:type → http://xmlns.com/foaf/0.1/Person

http://datos.bne.es/resource/XX1718747

rdf:type → http://.../Street

http://www.server2.es/resource/Cervantes

Author

D. Quijote

Date of Birth

1547

rdf:type → http://.../Municipality

http://geo.linkeddata.es/page/resource/Municipio/Cervantes

POLITÉCNICA

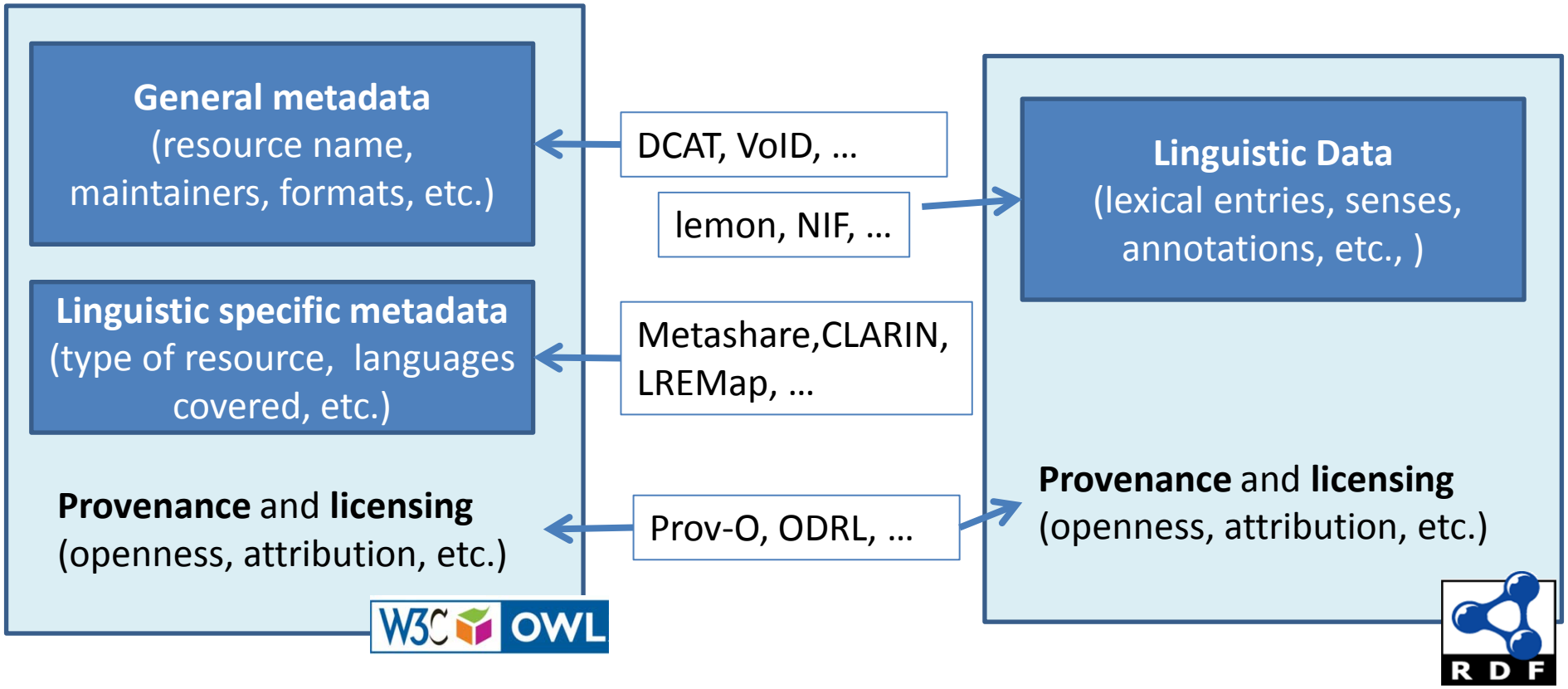
Linking LR content

- Which “sameAs” should I use?
  - myOwn: SameAs
  - lvont: somewhatSameAs
  - lvont: nearlySameAs
  - SKOS: exactMatch
  - SKOS: closedMatch
  - SKOS: relatedTo
  - OWL: sameAs

- Should be useful to introduce a confident measure in the link?

Linked Data allows linguistic metadata and linguistic data discovery, sharing, reuse and integration

Vocabularies for discovery Language Resources



Join the LD4LT W3C community group
http://www.w3.org/community/ld4lt/

A roadmap

1. Definition of the metadata OWL ontology @ LD4LT W3C group
   - Open community group
   - Bottom-up approach: UPF's model as starting point
   - Expanded with data and process PROVENANCE and LICENSE modules
   - Backwards compatible with MS and LREMap models
   - In agreement with members of LD4LT W3C group
2. Development of a generic RDF convertor of MS metadata
3. Exposure of metadata of MS nodes as LD
4. Develop a Linguistic LD observatory (LIDER)
5. In parallel, explore exposing data of MS LRs into LD
   - Definition of guidelines aligned with BPMLOD W3C group activities

MS License description as RDF

- META-SHARE recommends a set of 21 licenses, classified in:
  - META-SHARE Non Redistribution
  - META-SHARE Commons («distribution towards META-SHARE members»)
  - Creative Commons
- ALL of them can be represented as RDF using extendedly used vocabularies such as ODRL (Open Digital Rights Licenses)
- Advantages of expressing licenses as RDF:
  - Unambiguous identification of well known licenses by their URIs
  - Enables conditional access to resources
  - Automatic license compatibility analysis when integrating resources

Example of a MS License as RDF

META-SHARE NonCommercial NoRedistribution NoDerivatives For-a-Fee Licence

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix odrl:  <http://www.w3.org/ns/odrl/2/> .

<http://example.com/nc-nored-nd-ff> a odrl:Policy ;
    rdfs:label      "NC-NoReD-ND-FF" ;
    rdfs:comment    "MetaShare NonCommercial, No Redistribution, No Derivatives, for a fee.
                     Perpetual, worldwide, allowing no redistribution of the original. "@en ;
    rdfs:seeAlso    <http://www.meta-net.eu/meta-share....pdf> ;
    odrl:permission  [ a            odrl:Permission ;
                       odrl:action  odrl:reproduce;
                       odrl:duty    [ a            odrl:Duty ;
                                      odrl:action  odrl:pay ;
                                      odrl:target  "XXX EUR"
                                    ]
                     ] ;
    odrl:prohibition  [ a           odrl:Prohibition ;
                        odrl:action odrl:commercialize, odrl:distribute, odrl:derive
                      ] .
```

See demo at http://conditional.linkeddata.es

Conclussion. Linked Data is to be accessed and used by machines

Web channels

www.lider-project.eu

Join the community

www.w3c.org/community/ld4lt

twitter.com/multilingweb

Hashtag: **#LiderEU**